第 38 卷第 5 期
2010 年 5 月

同济大学学报(自然科学版)
JOURNAL OF TONGJI UNIVERSITY(NATURAL SCIENCE)

Vol. 38 No. 5
May 2010

文章编号：0253-374X(2010)05-0736-08

DOI:10.3969/j.issn.0253-374x.2010.05.019

基于本体的 VSM 在兴趣型学习社区分组中的应用

程　艳[1,2]，许维胜[1]，赵　斐[3]，何一文[1]

(1.同济大学 电子与信息工程学院，上海 201804；2.江西师范大学 计算机信息工程学院，江西 南昌 330022；
3.同济大学 经济与管理学院，上海 201804)

摘要：采用语义网络技术，提出了基于本体的向量空间模型(VSM)，计算学习者的兴趣向量，克服了传统的 VSM 有术语间语义相关性被忽略的不足，提高了兴趣相似性比较的精确程度，同时提出了一种基于学习者兴趣相似匹配度和学习者兴趣匹配浓度的学习社区自组织分组算法.针对模型使用本体中的概念构造向量空间表现出的巨大维数，运用概念索引降维法对兴趣特征矩阵进行合理降维，大大降低了计算的复杂性.最后通过应用案例验证分析了该模型算法具有较高的分组效率和良好的扩展性.

关键词：分组算法；本体；兴趣特征；向量空间模型；概念索引法

中图分类号：TP 391　　**文献标识码**：A

Application of Ontology-based VSM on Learning Community Construction of Interest

CHENG Yan[1,2], *XU Weisheng*[1], *ZHAO Fei*[3], *HE Yiwen*[1]

(1. College of Electronics and Information Engineering, Tongji University, Shanghai 201804, China; 2. College of Computer Information Engineering, Jiangxi Normal University, Nanchang 330022, China; 3. College of Economics and Management, Tongji University, Shanghai 201804, China)

Abstract: To get rid of the disadvantages caused by neglecting the semantic relevance between terms in the traditional vector space model, ontology-based vector space model(VSM) based on semantic web technology is presented to calculate the learners' interest eigenvector, which can enhance the relative accuracy of the interest similarity. And a self-organization grouping algorithm for community is put forward on the basis of the learners' interest similarity match-degree and its concentration. Great dimensions would take place with the ontology to construct vector space, thus concept indexing(CI) method and reasonable treatment to matrix of interest Eigen value are used to promote the calculation efficiency. Finally, an experimental analysis of online education cases is carried out to verify the model algorithm with high efficiency and good scalability.

Key words: grouping algorithm; ontology; interest eigenvector; vector space model; concept indexing method

目前国内外对教育虚拟社区的研究主要集中在基础研究，探索虚拟社区的基本原理，包括虚拟社区的概念、定义、原理和模式，如 Hope N. Tillman[1] 提出了教育虚拟社区的定义、特征、类型、社区服务、创建策略和交往发生等问题；Amy Jokim 博士[2] 于 2000 年提出建立社区的九大策略.而对虚拟社区技术发展研究以支撑虚拟社区的成长，包括在虚拟社区使用的工具及其技术潜力，探讨如何构建虚拟社区方面的研究较少.Sun Chuentsai，Lin S S J 采用基于距离的 RMHC(Random Mutation Hill Climbing)算法寻找一群给定学生的最优组划分的研究[3].A. Inaba，T. Tamura 和 R. Ohkubo 等开发的 TGF 系统，研究了从学习目标出发建立协作学习组[4-5].黄荣怀、林凉在 WebCLTM 协作学习平台中利用 Agent 构建的 E-Tutor 为系统提供分组功能[6].杨帆、申瑞民等提出一种基于信任奖励和动态交换机制的协作式自组织学习社区算法[7]等.当前的分组研究存在的主要问题有：组的概念并不能满足 CSCL 的要求，分组方式多是硬性分组，缺乏具有智能特性的组员选择算法，分组方法缺乏一定的智能性.

每位网络学习者如何能根据自己的兴趣信息进

收稿日期：2009-03-16
基金项目：国家自然科学基金资助项目(70871091，60804042)
作者简介：程　艳(1976—)，女，副教授，博士生，主要研究方向为人工智能、智能计算机辅助教育和智能控制等.
E-mail:chyan88888@163.com
许维胜(1966—)，男，教授，博士生导师，主要研究方向为自动化系统、智能控制和控制理论与应用等.
E-mail:icslab2@tongji.edu.cn

入最适合自己的兴趣社区小组进行协作式学习,是研究的重点.网络学习社区建立的重点和难点在于学习者之间相似关系的判定和计算.针对当前分组研究存在的不足,采用语义网络技术,提出了基于本体的向量空间模型(VSM),来计算学习者的兴趣向量,它克服了传统的 VSM 有术语间语义相关性被忽略的不足,提高了兴趣相似性比较的精确程度和智能性,将兴趣的隐性表示有效准确地显性表示出来,并根据兴趣相似度的思想,提出了一种基于学习者兴趣相似匹配度和相似匹配浓度的学习社区自组织分组算法.

1　基于本体的向量空间模型

向量空间模型(VSM)在文本处理、文本挖掘领域占有重要的地位.许多搜索引擎都使用这种模型来分类网络文档.

1.1　模型的形式化定义

定义 1　网络上获得的信息(网络文档、段落、句子、自然语言查询语句等)组成信息对象集合 $D=\{d_i \mid 1\leqslant i\leqslant M\}$,$M$ 为信息对象的总数.根据 VSM,每条信息 d_i 都可以用一个特征向量 $\boldsymbol{v}(s)=[s_1,s_2,\cdots,s_N]$ 来表示.s_i 与本体中的概念 c_i 对应,表示某个信息对象中术语 c_i 的权重.所有的信息对象表示成的特征向量就构成了向量空间 $\boldsymbol{V}=\{\boldsymbol{v}_1(s),\boldsymbol{v}_2(s),\cdots,\boldsymbol{v}_M(s)\}$.向量空间的维数等于本体构成的知识论域空间的维数 N.

1.2　语义相关度的计算

1.2.1　本体和有限本体图

本体表示的是结构化的知识,具有自然的层次结构,从图论的角度,本体可以表示成一张图.称为本体图.

定义 2　本体图 $G=\{V,\boldsymbol{E}_{dg}\}$,节点集合 V 对应于本体中实体的集合 $\{c_1,c_2,\cdots,c_N\}$,并且节点与实体存在着一对一的映射关系:$v_i\in V\leftrightarrow c_i$,$i=1,2,\cdots,N$,因此可以使用 $V(c_i)$ 表示图的节点 v_i.边 $e_{dg}=(u,v)$ 连接节点 $u\in V$ 和节点 $v\in V$,边集合 $\boldsymbol{E}_{dg}$ 对应于本体中的关系集合 $R:e_{dg}\leftrightarrow r_i\in R$,$i=1,2,\cdots,N$,因此可以用 $e_{dg}(V(c_i),V(c_j))$ 表示图中的边.

通过对本体图进行约束,例如对关系的约束,获得有限本体图.通过适当的约束,能够抛开次要问题,集中解决主要问题.基于本体 VSM 中的术语对应于本体中的概念,本体中概念的语义相关性度量通过基于有限本体图的语义距离计算获得.有限本体图是有向无环图,所以一定是拓扑有序,存在根节点 root,表现出层次结构.所以按照层次遍历的“Breadth-First”算法(类似于树图中的广度优先的遍历方法),扫描有限本体图,建立本体中的概念索引表.概念索引表层次关系明显,高层次的概念索引号在低层次的概念索引号的前面,并且相连的兄弟概念排列在一起,利用这样的特征能够简化语义距离的计算方法,不需多次扫描概念图,便于基于本体的自然语言文本的处理.

1.2.2　语义距离和语义相关度

语义距离是一种度量对象(概念、词汇和句子)间语义相似程度的概念,其数值表示形式通常是[0,∞]间的实数.语义距离与语义相关度成反比关系,例如两个概念间的语义距离越大,语义相关度就越低,反之,距离越小,相关度越大.语义相关度是一个强主观性的概念,与具体的应用相关,因此也很难获得一个统一的定义.考虑到本体可以表示成有限本体图,因此基于图形最短路径法来计算语义距离.

有限本体图上的节点与本体中的类/概念一一对应,对于本体中概念的集合 $O_c=\{c_1,c_2,\cdots,c_n\}$,语义距离函数 $d:O_c\times O_c\rightarrow R$ 是一个二元映射,并且满足:①非负性:$d(c_x,c_y)\geqslant 0$;②同一性:$d(c_x,c_y)=0$ 当且仅当 $c_x\equiv c_y$;③对称性:$d(c_x,c_y)=d(c_y,c_x)$;④三角不等性:$d(c_x,c_y)\leqslant d(c_x,c_z)+d(c_z,c_y)$.所以语义距离空间构成度量空间,称 d 是 O_c 上的度量函数记为 (O_c,d).通过语义距离矩阵描述该度量空间.

定义 3　对于有限本体图上的任意节点 $V(c_x)$ 到其自身的语义距离为 0,$d(V(c_x),V(c_x))=0$.

定义 4　对于有限本体图上的任意节点 $V(c_x)$ 和节点 $V(c_y)$,如果存在通路,那么定义语义距离为两节点间的最短路径长度 $d(V(c_x),V(c_y))=\min(V(c_x),V(c_y))$.

定义 5　对于有限本体图上的任意节点 $V(c_x)$ 和节点 $V(c_y)$,存在直接的连接(即 $\exists e_{dg}(V(c_x),V(c_y))$),如果两者间的关系是父子关系,那么定义语义距离为 $d(V(c_x),V(c_y))=1$;如果两者间的关系是对象属性关系,那么定义语义距离为 $d(V(c_x),V(c_y))=2$.

根据基于图形最短路径法计算语义距离,将本体概念间的语义距离用矩阵表示,获得语义距离矩阵为

$$\mathrm{dis}\boldsymbol{M}(O_c)=\begin{bmatrix} 0 & d(c_1,c_2) & \cdots & d(c_1,c_N) \\ d(c_2,c_1) & 0 & \cdots & d(c_2,c_N) \\ & & \cdots & \\ d(c_N,c_1) & d(c_N,c_2) & \cdots & 0 \end{bmatrix} \tag{1}$$

语义距离矩阵的维数 N 对应于本体中的类概念数,元素 $d(c_i,c_j)$表示术语 c_i 到术语 c_j 的语义距离.

给出由语义距离计算语义相关度的公式,$r(c_i,c_j)=e^{-\alpha d(c_i,c_j)}$,其中 $d(c_i,c_j)$对应于公式(1)中概念 c_i 和 c_j 间的语义距离值,α 表示陡度系数.此公式满足语义距离与语义相关程度间的对应关系:两个对象语义距离为0时,其语义相关度为1;两个对象语义距离为无穷大时,其语义相关度为0;两个对象语义距离越大,其语义相关度越小(单调下降).

根据语义相关度的计算公式,可以将语义距离矩阵 $\mathrm{dis}\boldsymbol{M}(O_C)$转化成语义相关矩阵 $\boldsymbol{R}(O_c)$.

$$\boldsymbol{R}(O_C)=\begin{bmatrix} 1 & r(c_1,c_2) & \cdots & r(c_1,c_N) \\ r(c_2,c_1) & 1 & \cdots & r(c_2,c_N) \\ & & \cdots & \\ r(c_N,c_1) & r(c_N,c_2) & \cdots & 1 \end{bmatrix} \tag{2}$$

2 基于本体 VSM 的兴趣特征量化匹配模型

提出基于本体的 VSM 计算学习者的兴趣特征向量,设计了基于本体 VSM 的学习者兴趣特征向量量化匹配模型,是一种考虑了术语间语义相关性的计算模型,根据学习者描述的兴趣信息,以学习者共同的知识背景——本体为基础,经过预处理和特征量化匹配两个功能模块处理,将文本信息表示成统一尺度的特征向量后,采用一定的方法,如余弦法,来进行学习者之间的兴趣相似匹配度计算.

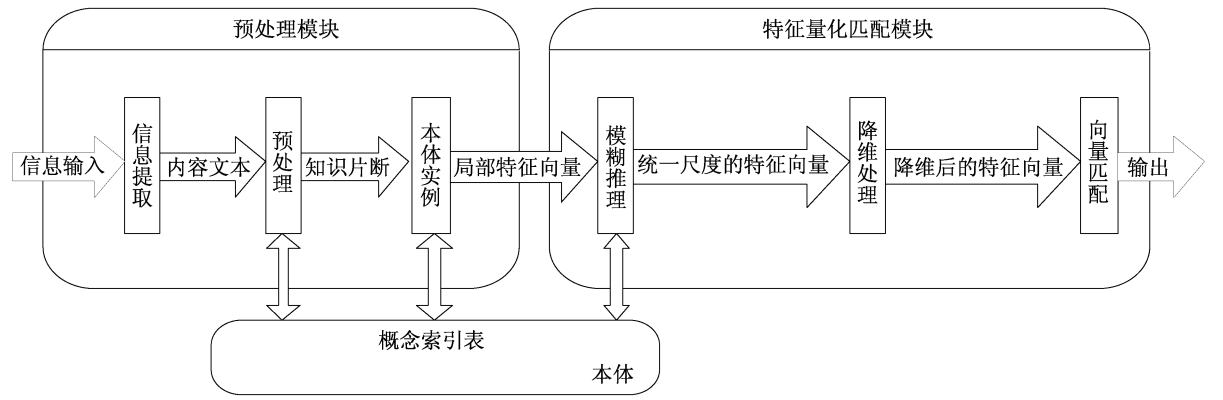


图1 特征量化匹配模型设计框图

Fig.1 Quantifying eigenvalue and matching model design block diagram

2.1 预处理模块

预处理模块主要功能是从描述信息中将属于知识论域中的术语提取出来,重构成本体实例.重构本体实例的具体步骤:①以本体的概念作为向量空间的术语项,对信息项内容中包含的术语进行提取.如果一个提取的概念在本体中找不到对应项,但却是本体中某些复合词的一部分,那么将所有相关的复合词都添加进来.②将提取的概念组成局部概念图,局部概念图表示的是用户的知识片断,作为本体的实例,它具有与有限本体图一致的特征,即为有向无环图.重构本体实例依从的规则为:以整条信息项所涵盖的知识集合为根节点 root;提取的概念间的相对关系与本体图中的相对关系保持一致,即上下位关系不变,旁支关系不变.

定义6 基于本体的 VSM,用户特征局部向量 $\boldsymbol{v}(s)=[s_1,s_2,\cdots,s_N]$,其中 $s_i(s_i\in[0,1])$是与本体中的 c_i 相对应.s_i 可描述为

$$s_i=\begin{cases} e^{-\alpha\mathrm{dis}(e_i,\mathrm{root})} \\ 0 \end{cases} \tag{3}$$

其中,如果对应的 c_i 出现在概念图中,则与之对应的 s_i 为 $e^{-\alpha\mathrm{dis}(e_i,\mathrm{root})}$,否则为0值.式(3)中的 dis($e_i$,root)表示局部概念图中概念 $e_i\in O_c$ 到 root 的语义距离,其计算方法也是基于图形的最短路径的计算方法;α 表示陡度估量,因为 $e^{-7}\approx 0$,在模糊建模中通常选择其值为 $-7/\max(\mathrm{dis})$.

2.2 信息的度量统一

从局部知识图(本体实例图)可知因为学习者用户用语的随意性,以及主观认定性,导致局部知识的

术语关系是依赖用户的相对关系,由此获得的特征向量是相对于个体局部知识结构的量化.另一方面,同一领域的概念间并不是孤立的,相互间具有一定的语义相关性,用户表明的本体实例中使用的概念往往隐含了与之相连的概念的信息.为了能够统一度量分散的特征向量,需要将局部知识转化到统一度量的空间中.本文采用模糊推理,基于概念间的相关性,将局部特征向量转化为统一度量尺度下的特征向量,以便后续的匹配比较.计算公式为

$$\boldsymbol{V}(s)\circ\boldsymbol{R}(o)\Rightarrow\boldsymbol{V}(s) \tag{4}$$

其中,$\boldsymbol{V}(s)=[s_1,s_2,\cdots,s_N]$表示局部知识的特征向量,根据式(3)计算,模糊关系矩阵 $\boldsymbol{R}(o)$已通过式(2)计算获得.$\boldsymbol{R}(o)$中每个元素 r_{ij}反映了术语间的相近程度.这样,那些在兴趣描述中没有显式地出现过的术语的语义关系也将被考虑.公式中的"∘"表示模糊关系中的内积,代表 max - min 操作.经模糊推理将分散的特征向量转化为统一度量空间中的特征向量.

$$s_i=\max(\min(s_1,r(1,i)),\min(s_2,r(2,i)),\cdots,\min(s_N,r(N,i))) \tag{5}$$

式(5)计算获得的 s_i 是统一度量空间中特征向量 $\boldsymbol{v}(s)$的元素.通过计算,来自学习者的个人兴趣描述信息都被转化成考虑语义后统一度量的特征向量.对于具有统一尺度的特征向量,可以进行比较匹配.匹配功能可以通过对语义特征向量进行计算相似度实现.参考向量空间模型中,目前向量相似性函数的计算方法包括内积法(Cosine 函数)、最近邻法(Max)、Minkowski 距离和 Euclidean 距离等.其中最直观有效的方法为内积法.

将用户 1 的特征向量表示为空间向量 $\boldsymbol{v}(d_1)=\{x_{11},x_{12},\cdots,x_{1n}\}$,将用户 2 的特征向量表示为空间向量 $\boldsymbol{v}(d_2)=\{x_{21},x_{22},\cdots,x_{1n}\}$.那么通过计算两向量 $\boldsymbol{v}(d_1)$和 $\boldsymbol{v}(d_2)$的夹角(内积)来度量向量间的接近程度.内积越大,两向量间的夹角越小.相似度量的计算公式为

$$S_{\text{sim}}(d_1,d_2)=\frac{\boldsymbol{v}(d_1)\cdot\boldsymbol{v}(d_2)}{|\boldsymbol{v}(d_1)||\boldsymbol{v}(d_2)|}$$

本文基于内积法(Cosine)来计算特征向量的相似性.代表用户 i 的兴趣特征向量 $\boldsymbol{v}_i=[s_{i1},s_{i2},\cdots,s_{iN}]$,代表用户 j 的兴趣特征向量 $\boldsymbol{v}_j=[s_{j1},s_{j2},\cdots,s_{jN}]$,首先根据 Cosine 法计算向量间的匹配度为

$$M_d(e(i,j))=\frac{|\boldsymbol{v}_i\wedge\boldsymbol{v}_j|}{|\boldsymbol{v}_i|\cdot|\boldsymbol{v}_j|} \tag{6}$$

通过计算两个向量 $\boldsymbol{v}_i$ 和 $\boldsymbol{v}_j$ 的内积来度量向量间的接近程度,内积越大,两个向量间的夹角越小,两个兴趣越接近.

2.3　概念索引(CI)降维

基于本体的信息提取系统,采用 VSM 来表达兴趣信息的文本特征,构造向量空间的维数依赖于本体中的概念数,表现出巨大的维数,导致"维数灾难",为此,需要先对特征矩阵进行合理的降维处理.现有的解决方法包括随机映射(RP),隐含语义分析(LSA),概念索引(CI)降维等等[8-10].其中概念索引(CI)是一种有效的降维方法,其速度与准确度综合效率较高.

采用 $\boldsymbol{v}(s)=[s_1,s_2,\cdots,s_N]$表示学习者用户兴趣信息的局部知识特征向量,设 $\boldsymbol{W}$ 为 $M\times N$ 矩阵.其中,M 为用户的数目,N 为本体构成的知识论域空间的维数;$v_1(s)=[s_{11},s_{12},\cdots,s_{1N}]$,$\boldsymbol{W}=[v_1(s),v_2(s),\cdots,v_M(s)]'$.矩阵 $\boldsymbol{W}$ 的第 i 行为第 i 个用户的兴趣向量空间表示 W_i,再设 r 为要降至的维数.先采用某种简单的聚类算法(K-means 或层次算法等)对矩阵 W 作 r 路聚类,将特征向量集分成 r 个互不相交的子集 $S1,S2,\cdots,Sr$,然后,对每个集合 Si 分别计算质心点向量 $\boldsymbol{C}_i$,并将它们规格化为单位向量 $\boldsymbol{C}'_i$,将每一个单位质心向量 $\boldsymbol{C}'_i$ 作为降维空间的一个坐标轴,形成一个 r 维子空间,每个用户的 r 维向量表示可通过将其映射到这个 r 维子空间得到,因此,降维后的兴趣向量空间 W'可通过式(7)的矩阵运算得到

$$\boldsymbol{W}'_{m\times r}=\boldsymbol{W}_{m\times n}\times\boldsymbol{C}_{n\times r} \tag{7}$$

其中,$\boldsymbol{C}_{n\times r}=(\boldsymbol{C}'_1,\boldsymbol{C}'_2,\cdots,\boldsymbol{C}'_r)$.$W'_{m\times r}$ 为降维后的用户兴趣局部特征向量构成的矩阵,第 i 行表示第 i 个用户的兴趣局部特征向量,局部特征向量由原来的 N 维降到 r 维.可以大大简化后面的计算量.

3　社区自组织分组算法

社区应该是一组具有相同或相似兴趣学习者组成的团体,因此社区的建立应该尽可能的将兴趣相似度大的学习者放在同一个社区中,因此提出如下的社区建立流程:

(1) 兴趣特征向量降维后,根据式(6)计算两两学习者之间的兴趣匹配度.

(2) 设定阈值 T_1,计算每个学习者的匹配浓度(由与该学习者兴趣匹配的学习者数目的多少来决定浓度的高低).假设某高校网络学习者为 m 个,学习者 i 的匹配浓度为 θ_i,计算公式为

$$\theta_i = \sum_{i=1}^{m} \alpha_{ij}/m$$

其中，$\alpha_{ij} = \begin{cases} 1 & M_d(e(i,j)) \geqslant T_1 \\ 0 & \text{其他} \end{cases}$ T_1 为预先假定的阈值.

(3) 选出匹配浓度最高的学习者，作为社区中心学习者来建立社区. 预先设定一个阈值 T_2，与中心学习者的兴趣匹配度高于阈值的学习者进入该社区.

(4) 对剩下的学习者按照流程(1)～(3)的顺序，进行重新计算，直到学习者的最高浓度低于一个阈值 T_0 为止. 对于浓度低于 T_0 的学习者，计算他与各社区中心学习者的相似匹配度，进入最相近的社区学习.

4 兴趣社区的应用案例分析

提出了一种基于本体兴趣向量空间模型的学习社区构建算法，将具有相似兴趣爱好的学习者自动有效地组织成一个个的学习者社区，帮助其共享资源，进行协作式学习. 实验结果证明，该算法具有较高的分组效率和良好的可扩展性.

4.1 算法实验步骤描述

(1) 给定领域本体图，采用语义距离的计算方法，获得本体中概念间的语义关系. 基于专家支持建立程序员研发的领域本体，为社区网络中分布的学习者用户提供共享的知识背景. 生成的本体概念关系如图2所示.

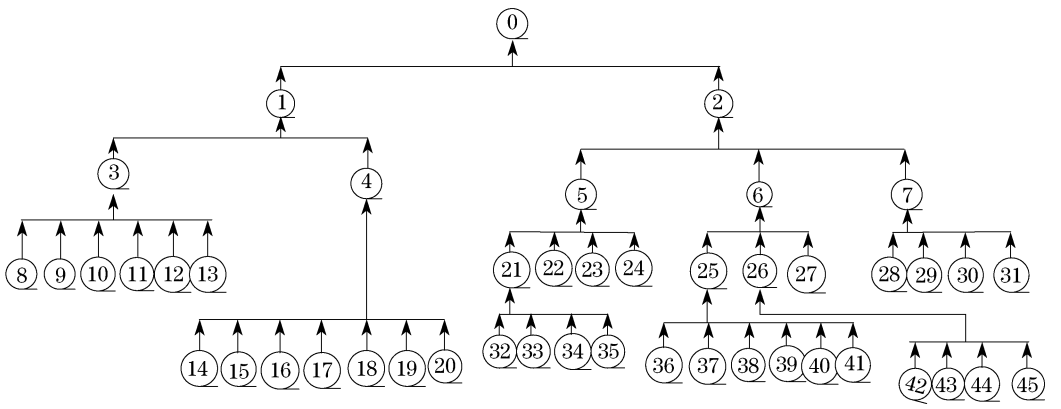


图2 程序员研发领域本体图

Fig.2 Programmer R & D domain ontology graph

从图2可以清晰地看出，知识可以分等级地表示，按照层次遍历的“Breadth-First”算法进行解析，与本体图2对应的概念索引表为

{0 Computer Programmer-Related-Technology, |1 NETWORK, 2 software, |3 Network Protocol, 4 WEB Development, 5 Operation-system, 6 Programming-language, 7 database, |8 SNMP, 9 ATM, 10 DSN, 11 APR/RAPR, 12 UDP, 13 TCP/IP, 14 HTML, 15 PHP, 16 JSP, 17 Ruby, 18 Ajax, 19 Java Script, 20 UI, 21 Embedded-operation-system, 22 Macintosh-operation-system, 23 windows, 24 DOS, 25 Developer, 26 .NET, 27 Assembly-language, 28 Oracle, 29 DB2, 30 MS-SQL, 31 MySQL, |32 WinCE, 33 VxWorks, 34 ucLinux, 35 Linux, 36 C, 37 MFC, 38 VB, 39 C++, 40 JAVA, 41 Delphi, 42 ASP.NET, 43 C#, 44 VB.NET, 45 VC.NET}.

索引列表中的竖直线用来示意有限本体图中的概念的层次划分. 这里，概念对应的标号并非随机指定的，而是与概念所在的层次排列密切相关，图中的标号对应于概念索引的序列号. 考虑到本体可以表示成有限本体图，那么基于图形最短路径法来计算语义距离. 根据语义距离矩阵的定义及式(1)，获得如图3所示的 46×46 语义距离矩阵.

从图3可以看出，语义距离矩阵是个对称矩阵，灰度度量了实体间语义距离的大小. X，Y 轴的标号对应于本体概念索引表中的标号.

(2) 网络学习者兴趣描述信息的局部知识数值化，并经过模糊推理转化成统一度量下的特征向量.

假如学习者甲描述自己的兴趣信息项为“Software Engineer, experiences in JAVA language, interested in WEB development with Ruby or HTML”，依据前面重构本体实例的方法从中提取出

的知识空间内的术语为“Software, JAVA, WEB development, Ruby, HTML”,根据术语在知识空间内的相对结构,上面信息中提取的知识片断可以表示成概念图 4.学习者乙描述的兴趣信息项提取出的知识空间内术语为“Operation System, Embedded_Operation_System, Linux, Database, MySQL”,同样的方法可以解析成本体实例,转化成概念图 5.

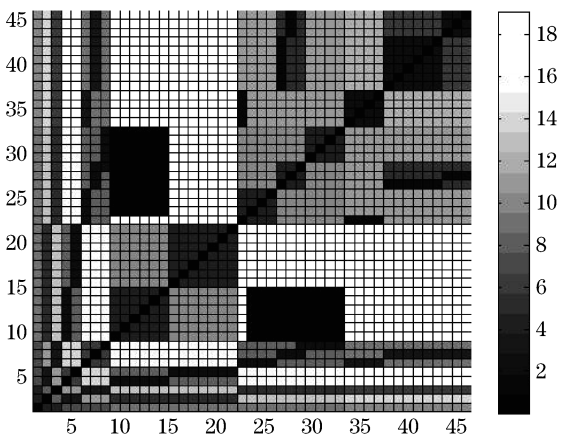


图 3　程序员研发本体的语义距离矩阵示意图

Fig.3　Semantic distance matrix diagram of programmer R & D domain ontology

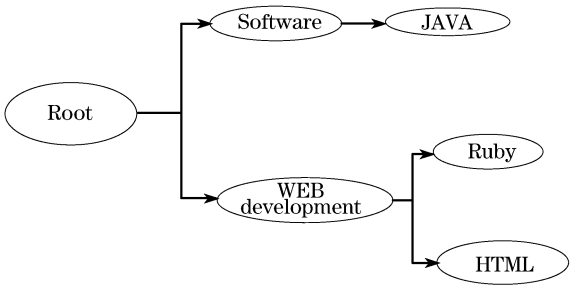


图 4　学习者甲的本体实例的概念图

Fig.4　Concept map of learner A's ontology graph

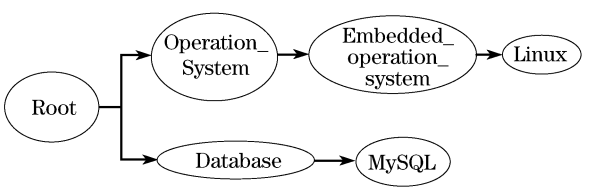


图 5　学习者乙的本体实例的概念图

Fig.5　Concept map of learner B's ontology graph

根据定义 6 和式(3),甲、乙学习者兴趣描述信息的局部知识数值化为兴趣特征局部向量为

$$\boldsymbol{v}_1(s):[0,0,e^{-\alpha},0,e^{-\alpha},0,\cdots,e^{-2\alpha},0,\cdots,e^{-2\alpha},0,\cdots,e^{-2\alpha},0,0,0,0,0]$$

$$\boldsymbol{v}_2(s):[0,0,0,0,0,e^{-\alpha},0,e^{-\alpha},0,\cdots,e^{-2\alpha},0,\cdots,e^{-2\alpha},0,0,0,e^{-3\alpha},0,\cdots]$$

特征向量中的元素对应于概念索引表中的项.现假设有 $M=12$ 个学习者希望能够通过自己声明的兴趣,为其发现适合自己的学习社区.同样的方法,从网络学习者的兴趣信息描述中提取出兴趣特征术语,并对其进行重构,由此获得的其他学习者用户兴趣描述信息项的局部知识向量为

$$\boldsymbol{v}_3(s):[0,0,e^{-\alpha},0,e^{-\alpha},0,\cdots,e^{-2\alpha},0,\cdots,e^{-2\alpha},0,\cdots,e^{-2\alpha},0,\cdots,e^{-3\alpha},0,e^{-3\alpha}]$$

$$\boldsymbol{v}_4(s):[0,0,0,e^{-\alpha},e^{-\alpha},0,\cdots,e^{-2\alpha},0,\cdots,e^{-3\alpha},0,\cdots,0]$$

$$\boldsymbol{v}_5(s):[0,0,e^{-\alpha},0,\cdots,0,e^{-\alpha},e^{-\alpha},0,\cdots,e^{-2\alpha},e^{-2\alpha},0,\cdots,0]$$

……

(3) 经过模糊推理,学习者的局部知识被转化成统一度量下的特征向量.然后根据用户设置的维数 r 按照 2.3 采用 CI 法计算降维的兴趣特征向量.

将每个学生的兴趣特征向量 $\boldsymbol{v}_i(s)$ 视为 $N=46$ 维的向量(N 为本体构成的知识论域空间的维数),$M=12$ 个学生的兴趣局部知识向量可以表示成 $M\times N$ 矩阵 $\boldsymbol{W}$,$\boldsymbol{W}_{12\times 46}=[\boldsymbol{v}_1(s),\boldsymbol{v}_2(s),\cdots,\boldsymbol{v}_{12}(s)]'=\begin{bmatrix}\boldsymbol{v}_1(s)\\ \boldsymbol{v}_2(s)\\ \vdots\\ \boldsymbol{v}_{12}(s)\end{bmatrix}$,假设用户设置的维数 $r=3$,将矩阵 W 简单分成 3 个互不相交的子集,

$$s_1=\{\boldsymbol{v}_1(s),\boldsymbol{v}_4(s),\boldsymbol{v}_7(s),\boldsymbol{v}_{10}(s)\},$$

$$s_2=\{\boldsymbol{v}_2(s),\boldsymbol{v}_5(s),\boldsymbol{v}_8(s),\boldsymbol{v}_{11}(s)\},$$

$$s_3=\{\boldsymbol{v}_3(s),\boldsymbol{v}_6(s),\boldsymbol{v}_9(s),\boldsymbol{v}_{12}(s)\}$$

对每个集合 s_i 分别计算得质心点向量 $\boldsymbol{C}_i$,得

$$\boldsymbol{C}_{46\times 3}=(\boldsymbol{C}_1',\boldsymbol{v}_2',\boldsymbol{v}_3')=\begin{bmatrix}0 & 1.553 & 0\\ 1.478 & 0 & 1.899\\ 1.324 & 1.678 & 1.972\\ \vdots & \vdots & \vdots\\ 1.285 & 1.561 & 0\end{bmatrix}$$

$\boldsymbol{W}'_{12\times 3}=\boldsymbol{W}'_{12\times 46}\times\boldsymbol{C}_{46\times 3}$.$\boldsymbol{W}'_{12\times 3}$ 为降维后的用户兴趣局部特征向量构成的矩阵,第 i 行表示第 i 个学习者的兴趣局部特征向量,局部特征向量由原来的 $N=46$ 维降到 $r=3$ 维.可以大大简化随后的计算量.

(4) 根据学习者兴趣特征向量间的相似度,自

组织学习社区.根据式(6)计算出 $M=12$ 个学生用户 user 之间的匹配程度,使用矩阵表示为

user1	1	0.229 33	0.229 33	0.229 33	0.229 33	0.229 33	0.229 33	0.696 47	0.643 77	0.653 92	0.468 11	0.229 33
user2	0.229 33	1	0.298 16	0.240 05	0.699 39	0.699 39	0.298 16	0.229 33	0.229 33	0.229 33	0.229 33	0.699 39
user3	0.229 33	0.298 16	1	0.240 05	0.298 16	0.298 16	0.578 77	0.229 33	0.229 33	0.229 33	0.229 33	0.298 16
user4	0.229 33	0.240 05	0.240 05	1	0.240 05	0.240 05	0.240 05	0.229 33	0.229 33	0.229 33	0.229 33	0.240 05
user5	0.229 33	0.699 39	0.298 16	0.240 05	1	0.700 74	0.298 16	0.229 33	0.229 33	0.229 33	0.229 33	0.700 74
user6	0.229 33	0.699 39	0.298 16	0.240 05	0.700 74	1	0.298 16	0.229 33	0.229 33	0.229 33	0.229 33	0.748 46
user7	0.229 33	0.298 16	0.578 77	0.240 05	0.298 16	0.298 16	1	0.229 33	0.229 33	0.229 33	0.229 33	0.298 16
user8	0.696 47	0.229 33	0.229 33	0.229 33	0.229 33	0.229 33	0.229 33	1	0.643 77	0.653 92	0.468 11	0.229 33
user9	0.643 77	0.229 33	0.229 33	0.229 33	0.229 33	0.229 33	0.229 33	0.643 77	1	0.643 77	0.468 11	0.229 33
user10	0.653 92	0.229 33	0.229 33	0.229 33	0.229 33	0.229 33	0.229 33	0.653 92	0.643 77	1	0.468 11	0.229 33
user11	0.468 11	0.229 33	0.229 33	0.229 33	0.229 33	0.229 33	0.229 33	0.468 11	0.468 11	0.468 11	1	0.229 33
user12	0.229 33	0.699 39	0.298 16	0.240 05	0.700 74	0.748 46	0.298 16	0.229 33	0.229 33	0.229 33	0.229 33	1

输入阈值0.3,在原来的模糊矩阵中,模糊系数大于给定阈值的将被置成1,小于阈值的将被置成0.

user1	1	0	0	0	0	0	0	1	1	1	1	0
user2	0	1	0	0	1	1	0	0	0	0	0	1
user3	0	0	1	0	0	0	1	0	0	0	0	0
user4	0	0	0	1	0	0	0	0	0	0	0	0
user5	0	1	0	0	1	1	0	0	0	0	0	1
user6	0	1	0	0	1	1	0	0	0	0	0	1
user7	0	0	1	0	0	0	1	0	0	0	0	0
user8	1	0	0	0	0	0	0	1	1	1	1	0
user9	1	0	0	0	0	0	0	1	1	1	1	0
user10	1	0	0	0	0	0	0	1	1	1	1	0
user11	1	0	0	0	0	0	0	1	1	1	1	0
user12	0	1	0	0	1	1	0	0	0	0	0	1

最后根据匹配浓度法自组织学习社区.user1,user8,user9,user10,user11 进入社区1学习;user2,user5,user6,user12 进入社区2;user3,user5,user7进入社区3;user4 进入社区4.

4.2 结果分析

(1) 特征向量降维时,设置的维数 r 越小,对应的概念数 N 越小,算法的复杂度就越小.没有采用CI特征降维设置时,默认情况下的计算复杂度依赖于本体中概念的数量,当用CI降维后,计算复杂度将直接受用户设置维数的影响.

(2) 根据学生用户之间的兴趣匹配度,设定阈值 T_1,计算每个学习者的匹配浓度,求出1值最多者(即浓度最高者),然后以此为中心组建社区.阈值过小,社区数较少,每个社区的学习者人数越多,会将兴趣不同的学习者归入同一社区.反之,阈值过大,会使社区数目增多,会使兴趣相近的学习都分到不同社区学习.故阈值的大小要根据经验取值,不宜过大或过小.

4.3 性能评价

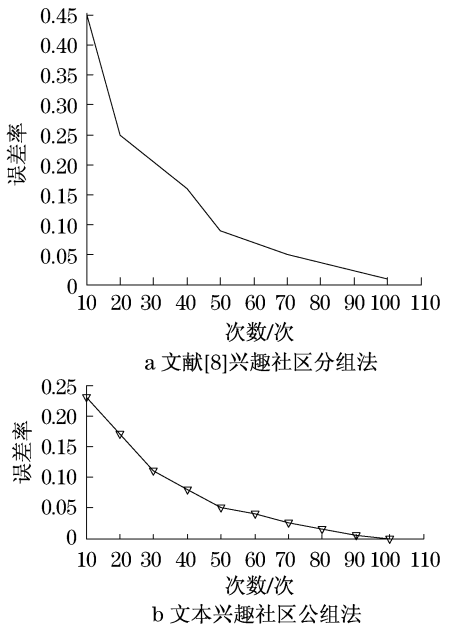
基于本体的VSM,考虑概念相关性,以同济大学网络教育学院网络学习学生为实验对象进行实验研究,实验表明,该兴趣社区自组织分组算法具有较高的分准率和分全率.定义分准率和分全率的计算公式为

$$\text{分准率} = \frac{\text{社区学习者认同数}}{\text{社区学习者总数}}$$

$$\text{分全率} = \frac{\text{社区学习者总数}}{\text{参与学习者总数}}$$

分准率的计算采用人工判断返回信息项的方法计算,例如对20名网络学习者进行社区划分后,人工判断与学习者用户期望一致得到肯定的信息项数为 n,那么划准率为 $p=n/20$,分别对学生人数 M 为20,50和100时进行了实验.在 $M=20$ 时,平均分准率=83%(误差率17%);$M=50$ 时,平均的分准率=95%(误差率5%);$M=100$ 时,平均的分准率=98%(误差率2%).实验表明,实验对象越多,分准率越高.对于算法中浓度低于 T_0 的学习者,计算他与各社区中心学习者的相似度,进入最相近的社区学习,故该算法分全率达到近100%.学生能根据自己描述的兴趣进入最适合自己的社区进行学习.杨帆、申瑞民[7]等提出的打分/交换的用户动态聚类算法,从学生的资源请求中发现学生兴趣,并有效地将具有相同兴趣的学生自动组成学习社区.在每个学生平均提出了20次请求后,系统分组的分准率为75%,在每个学生平均提出了50次请求后,系统分组的分准率提高到91%(即误差率由25%降到9%),在每个学生平均提出100次请求后,具有相同兴趣的学

生都聚集在同一个小组中，并趋于动态平衡，社区分组误差趋于零.两种方法分组误差比较见图 6.



a 文献[8]兴趣社区分组法

b 文本兴趣社区公组法

图 6　分组误差图

Fig.6　Grouping error convergence

由图可见，基于语义网络技术的兴趣社区自组织分组算法，随着加入社区学习者人数的增加，分组误差逐渐减少，当社区人数增加至 100 时，误差率趋于零.在初始阶段，杨帆方法的误差率很高，本文方法的分组准确度大大高于前者，随着学生请求次数的增多或加入社区学习的学生人数的增加，分组准确率大幅度增加，误差率大大减少，特别是实验参数值达到 100 时，误差率接近零，由图可知，两者分组误差都具有较好的收敛性，本方法具有更高的分准率.

5　结语

网络学习者社区是一组拥有相同或相似兴趣的学习者团体，每位网络学习者如何能根据自己描述的兴趣信息进入最适合自己兴趣的兴趣社区小组进行协作式学习，是本文研究的重点.基于本体的 VSM 计算学习者的兴趣特征向量，是一种考虑了术语间语义相关性的计算模型，根据学习者描述的兴趣信息，以学习者共同的知识背景——本体为基础，根据兴趣的隐性表示获取对应的显式表示（即兴趣向量）；提出了一种基于学习者兴趣相似匹配度和学习者相似匹配浓度的学习社区的自组织分组算法，是协作学习分组研究中一种优化分组方法；针对使用本体中的概念构造向量空间表现出的巨大维数，把概念索引降维法改进运用于兴趣特征降维，对兴趣特征矩阵进行合理的降维处理，大大降低了运算量和计算复杂度，提高了算法效率.

参考文献：

[1] Hope N. Tillman. Virtual community building using internet tools[EB/OL]. [2004 - 10 - 16]. http: // www. Hopetillman. com/i100/vc. htm

[2] Amy Jokim. Network community construction—design strategy unmasked[M]. Translated by ZHANG Shu, HU Rong, ZHAO Ming. Beijing: Tsinghua University Press, 2001.

[3] Sun C T, Lin S S J. Learning through collaborative design: a learning strategy on the internet[C/OL] // 31st ASEE/IEEE Frontiers in Education Conference. [2006 - 12 - 01]. http: // citeseer. nj. nec. com/505392. html.

[4] Inaba A, Tamura T, Ohkubo R, et al. Design and analysis of learners' interaction based on collaborative learning ontology [C] // Proceedings of Euro - CSCL2001. Maastricht: [s. n.], 2001: 308 - 315.

[5] Inaba A, Ikeda M, M Izoguch I R. Learning goals and design rationales in collaborative learning: an ontological approach to support design of collaborative learning[EB/OL]. [2006 - 12 - 12]. http://www. ei. sanken. osakau. ac. jp/pub/ina/isir03. pdf.

[6] 黄荣怀，林凉. 构建 WebCL 平台上的 e-Tutor[EB/OL]. [2006 - 12 - 20]. http: // www. etc. edu. cn/articledigest15/goujian. htm.
HUANG Ronghuai, LIN Liang. Building e-Tutor on WebCL platform[EB/OL]. [2006 - 12 - 20]. http: // www. etc. edu. cn/articledigest15/goujian. htm.

[7] 杨帆，申瑞民，童任，等. 一种新颖的协作式自组织学习社区算法[J]. 上海交通大学学报，2004，38(12)：2078.
YANG Fan, SHEN Ruimin, TONG Ren, et al. A novel collaborative self-organizing learner communities algorithm [J]. Journal of Shanghai Jiaotong University, 2004, 38 (12): 2078.

[8] Lin J, Gunopulos D. Dimensionality reduction by random projection and latent semantic indexing [C] // Text Mining Workshop, the 3rd SIAM International Conference on Data Mining. Chicago: University of Illinois. Army High Performance Computing Research Center, 2003: 301 - 309.

[9] Fodor I K. A survey of dimension reduction techniques [EB/OL]. [2008 - 12 - 20], http: // www. llnl. gov/CASC/sapphire/pubs. html, 2002.

[10] George Karypis, Han Euihong. Concept indexing: a fast dimensionality reduction algorithm with applications to document retrieval & categorization [C] // Proceedings of 9th International Conference on Information and Knowledge Management. McLean: ACM SIGIR, ACM SIGMI, 2000: 1 - 20.